Ensembles in Face Recognition: Tackling the extremes of high dimensionality, temporality, and variance in data

Nitesh V. Chawla
Dept. of Computer Science and Engineering
University of Notre Dame
IN 46556 USA
nchawla@cse.nd.edu

Kevin W. Bowyer
Dept. of Computer Science and Engineering
University of Notre Dame
IN 46556 USA
kwb@cse.nd.edu

Abstract

Random subspaces are a popular ensemble construction technique that improves the accuracy of weak classifiers. It has been shown, in different domains, that random subspaces combined with weak classifiers such as decision trees and nearest neighbor classifiers can provide an improvement in accuracy. In this paper, we apply the random subspace methodology to the 2-D face recognition task. The main goal of the paper is to see if the random subspace methodology can improve the performance of the face recognition system given the high dimensional data, temporal, and distribution variant data. We used two different datasets to evaluate the methodology. One dataset comprises of completely unique subjects for testing, and the other dataset comprises of the same subjects (both in training and testing) but images in the test set are captured at different times under different conditions.

1 Introduction

Face images are usually represented as high-dimensional pixel matrices, where in each matrix cell is a gray-level intensity value. These raw feature vectors can be very large and highly correlated. Moreover, the size of the enrollment data is usually small. This small sample size coupled with the very high-dimensionality of raw feature vectors can lead to a difficult pattern recognition task (*the curse of dimensionality*). The lack of enough samples in very high dimensions can reduce the accuracy of nearest neighbor classifiers [1, 2]. In addition, the extreme sized dimensions can present scalability issues.

To combat these issues of very high feature correlation, small sample size and computational complexity, the face images are often transformed into a lower dimensional manifold. One of the most popular techniques for linear transformation in feature space is PCA [3, 4, 5]. PCA reduces the dimensions by rotating feature vectors from a large highly correlated feature space *(image space)* to a smaller feature space *(face space)* that has no sample covariance between the features.

The face space is typically improved by a filtering phase, wherein some number of the highest and/or lowest eigen values are discarded. This further reduces the dimensionality and increases the stability of the classifier. But, there is no generally accepted procedure for the number of eigen values to drop from front or behind. There is some evidence that the initial face space dimensions might be lighting variations. But this is of course dependent on the conditions represented in the set of images. Some studies drop the highest eigen value and retain 60% of the remaining vectors [6]. Typically, the studies directly tune the performance on the testing set, and establish an operating point. However, that performance can be misleading as it is overfit on the testing set, and the generalization accuracy of the classifier cannot be sufficiently established.

The nearest neighbor classifier, a popular choice in the 2-D face-recognition domain, can be very sensitive to the sparsity in the high-dimensional space. Their accuracy is often far from optimal because of the lack of enough samples in the high-dimensional space [1, 2]. Bootstrapping is commonly applied to mitigate the issues with sparsity in data [7]. However, bootstrapping can only enrich the sample space, but not reduce the high-dimensionality, which can persist to be a problem. The random subspace method [8] can effectively exploit the high dimensionality of the data. The random subspace method constructs an ensemble of classifiers on independently selected feature subsets, and combines them using a heuristic such as majority voting, sum rule, etc.

In this paper, we evaluate random subspaces for countering the high-dimensionality of feature space, data sparsity, temporal and distribution variations in data. We include a component of our recent paper on random subspaces

for face recognition [9]. We use a nearest neighbor classifier with the Mahalanobis Cosine (MahCosine) distance measure in this setting. We chose MahCosine as it is relatively stable and has been shown to give better performances than other distance based measures [10]. Wang and Tang [11] previously used random subspaces for face recognition. Howevery, they pre-selected the top 50 dimensions and randomly selected the other 50. They noticed a performance drop if they chose all 100 at random. However, in our experiments we see that constructing completely random subspaces is as good if not better than a single classifier.

2 Classifier construction

In this section, we discuss in brief the PCA methodology and the MahCosine distance metric as implemented in the CSU code [10]. We start with the raw feature vectors for each image, apply PCA, and then apply the ensemble techniques of subsampling and random subspace. Depending on the experiment, we use either the tuned or the complete face space. We evaluate the various scenarios, as suggested in the framework, in this paper.

2.1 PCA

PCA performs a linear transformation in the raw feature space to construct a lower dimensional manifold [4]. The raw feature vectors are a concatenation of the gray-level pixel values from the images. Let us assume there are *m* images and *n* pixel values. Let *Z* be a matrix of *(m,n)*. The mean image of *Z* is then subtracted from each of the images in the training set, $\Delta Z_i = Z_i - E[Z_i]$. Let the matrix *M* represent the resulting "centered" images; $M = (\Delta Z_1, \Delta Z_2, ..\Delta Z_m)^T$. The covariance matrix can then be represented as: $\Omega = M.M^T$. Ω is symmetric and can be expressed in terms of the singular value decomposition $\Omega = U.\Lambda.U^T$, where U is an m x m unitary matrix and $\Lambda = diag(\lambda_1, ..., \lambda_m)$. The vectors $U_1, ..., U_m$ are a basis for the *m-dimensional* subspace. The covariance matrix can now be re-written as

$$\Omega = \sum_{i=1}^{m} \zeta_i.U_i$$

The coordinate ζ_i, $i \in 1, 2, ...m$, is called the i^{th} the principal component. It represents the projection of ΔZ onto the basis vector *U*. The basis vectors, U_i, are the principal components of the training set. Once subspace is constructed, the recognition is done by projecting the a centered probe image and gallery image into the subspace, and the closest gallery image to probe image is selected as the match.

2.2 Distance measures

Various distance measures have been evaluated in the realm of face recognition [12, 6]. For our experiments, we utilized the MahCosine distance metric [10]. Our initial experiments showed that MahCosine significantly outperformed the other distance measures.

The MahCosine measure is the cosine of the angle between the images after they have been transformed to the Mahalanobis space and normalized by the variance estimates [10]. Formally, the MahCosine measure between the images *i* and *j* with projections *a* and *b* in the Mahalanobis space is computed as:

$$MahCosine(i, j) = cos(\theta_{ij}) = \frac{|a||b|cos(\theta_{ij})}{|a||b|}$$

As a part of future work, we are also going to consider different distance measures and evaluate the improvements obtained by the random subspaces and resampling techniques.

2.3 Random subspaces

The random subspace method, introduced by Ho [8], randomly selects different feature dimensions and constructs multiple smaller subsets. A classifier is then constructed on each of those subsets, and a combination rule is applied in the end for prediction on the testing set. For the nearest neighbor algorithm, it simply means that only a randomly selected subset of the complete face space contributes towards the distance computation. The random subspace methodology can be potentially useful for face recognition due to the inherent sparsity and small-sample size of data. For each random subspace a different nearest neighbor classifier is constructed projecting the test feature vector into a different (but potentially) overlapping face space. Given an m x $(m-1)$ dimensional face space (for *m* training images, there are at most *m-1* non-zero eigen values), where *m* is the number of images, the feature vector can be represented as $X = (x_1, x_2, ..., x_{m-1})$. Then, multiple random subspaces of size *m x p* are selected, k times, where *p* is the size of the randomly selected subspace, $X_p^k = \{(x_1, x_2, ..., x_p)|p < (m-1)\}$.

For each subject in the testing set, the *1*-nearest neighbor is found among each of the *m-1* dimensional subspaces, using the procedure briefly outlined in the PCA discussion (Section 2.1). This process is repeated for a pre-selected *K* number of times. The classification can either be done by taking the most popular class attached to the test subject or by aggregating the distance measure computed from each of the subspaces. We aggregated the distance measure for our experiments.

The random space method can be outlined as follows:

1. For each *k=1,2,..K*

(a) Select a p dimensional random subspace, X_p^k, from X.

(b) Project the probe and gallery set onto the subspace X_p^k.

(c) Construct the nearest neighbor classifier, C_p^k, using the Mahalanobis Cosine metric. Compute the corresponding distances by C_p^k for each gallery and probe image.

2. Aggregate the distances assigned to the probe and gallery images by each of the C^k classifiers. The aggregation is a simple average of all the individual distances.

3. Rank order the images and compute the rank-one accuracy.

The individual classifiers can be weaker than the aggregate or even the global classifier. Moreover, the subspaces are sampled independently of each other. An aggregation of the same can lead to a reduction in the variance component of the error term, thereby reducing the overall error [13, 14]. There is a popular argument that diversity among the weak classifiers in an ensemble contributes to the success of the ensemble [15]. The diversity of classifiers in an ensemble is considered a key issue in the design of an ensemble. Classifiers are considered diverse if they disagree on the kind of errors they make. Diversity is an important aspect of the ensemble techniques — bagging, boosting, and randomization [16]. One can, for example, construct a correlation measure among the rank-orders provided by each of the individual classifiers to get an estimate of diversity.

In addition, the random subspace technique also counters the sparsity in the data, as the subspace dimensionality gets smaller but the training set size remains the same.

While the random subspace method has been applied in various machine learning tasks, it has not received significant attention in face recognition task. In addition to our recently accepted paper in CVPR [9], we are only aware of only a recent paper by Wang and Tang [11] for Face recognition. They applied a variant of the random subspace methodology using LDA as the base classifier. Their technique used ($N = 100$) dimensions, and always selected the largest $N_0 = 50$ dimensions. The remaining $N_1 = N - N_0$ dimensions were selected randomly from the rest of the face space. They note that by selecting N_1 dimensions randomly a certain element of diversity is maintained in the ensemble.

We consider the original random subspace methodology and use a nearest neighbor classifier. The goal is to see if completely random selection of the subspace performs as good, if not better, than a carefully selected tuned space. By introducing a pre-determined set of dimensions, it implies that a certain element of tuning is required. As we mentioned in the Introduction, tuning of the face space on the testing set leads to a biased estimate of the performance. Moreover, we wanted to evaluate the enemble approaches in the nearest neighbor setting.

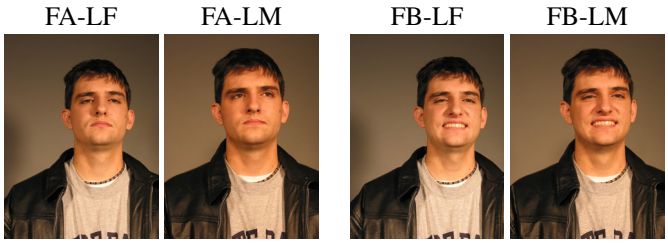

Figure 1. Sample images of a subject in the training data.

3 Data

The data for this paper was acquired from that available from the University of Notre Dame [17], and from the Feret database [18]. At Notre Dame, the subjects participate in the acquisition repeatedly (at most once in a week) over a period of time. The images of the subjects captured with two side lights on (*LF*) and two side lights and a center light on (*LM*). In addition the subjects are imaged with two expressions: neutral (*FA*) and smile (*FB*). The nomenclature is as used by FERET [18]. The data used in this paper was acquired during Spring 2002, Fall 2002, and Spring 2003. The color images in Spring 2002 were taken using a Sony MVC-95 camera with JPEG image sizes of 1600x1200. The images post-Spring 2002 were taken by the Canon Powershot G2 cameras, which provide 1600x1200 or 2272x1704-pixel images in JPEG format. Figure 1 shows sample images of a subject captured under the four conditions.

We considered two sets of data for this paper. In one, the probe and gallery sets comprised of subjects that were completely unique of the training set. In this one, the training set is comprised of 600 images of which 462 are from the FERET database and 138 are from the University of Notre Dame (ND) database. The earliest and latest images of another 393 subjects were partitioned into a gallery set and probe set. *Gallery* is the set of subjects enrolled in the database; and *Probe set* is the "testing" set. The goal is to project the probe set into the trained face space and correctly match it with the projected representative in the gallery. The earliest images of the 393 subjects were selected to create the gallery set, and the latest images were for the probe set. We only considered the **FA-LF** images. Please note that the gallery and probe images were all from the University of Notre Dame data. We will call this data as *Data-1* for subsequent discussions. This dataset exhibited the property of different distributions in the training and testing sets.

For the second set, we selected all the subjects from the

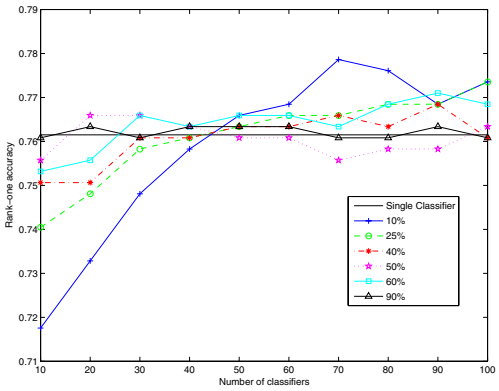

Figure 2. Results for *Data-1*.

Notre Dame repository that participated in at least three acquisition sessions. The first session was utilized as the training set, the second session became the gallery set, and the last session became the probe set. Thus, each set comprised of same subjects but different unique images. This gave us a total of 358 subjects with four images captured during each acquisition session: four different lighting and expression combinations: *FA-LF*, *FB-LF*, *FA-LM* and *FB-LM*. For training, we combined all the four images of each subject into a single set, giving a total of 1432 images. We constructed a face space from all the 1432 images. We then applied random subspaces to that face space. However, for testing purposes we report on each of the four sets individually. We wanted to evaluate the sensitivity of the random subspaces performance to the testing sets acquired in different conditions. We also wanted to examine if the random subspaces mitigate the need of knowing the meta-information (lighting condition, expression, etc.) about each image in the testing set. We will call this data as *Data-2*. This dataset exhibited the property of temporal and condition (expression and lighting) variations in data.

4 Experiments

We first elucidate our results using *Data-1*. It has a testing set of 393 subjects in each of the gallery and probe sets. Given the training set size of 600, the basis vector count is 599. For the random subspace experiments we randomly selected subsets of the following proportions: 10%, 25%, 40% 50%, and 90%. We compared the performances to the classifier constructed on the complete face space. We set the ensemble size to be 100 for random subspaces.

Figure 2 shows the result of using the random subspace methodology. We note that smaller sized subspaces start at a lower point, but get ahead of the larger sized subspaces. This is not a surprising result given the sparsity of the data. Ho [8] also noted a similar result for the hand-writing character recognition task. The classifiers with larger sized subspaces start at a higher point but then plateau. The weaker classifiers can potentially be generalizing better and exhibit a larger diversity in the ensemble. The goal is to reduce the variance component of the classifiers. By combining classifiers that are reasonably independent of each other, the error can be reduced. When the subspace size is very large, the classifiers are not very diverse. This lack of variance affects the overall performance of the ensemble. However, all the different subspace sizes tend to exceed or equal the rank-one accuracy obtained by the classifier learned on the complete face space.

Figure 3 shows the box-plots for the base classifiers used in the random subspaces. The box plot shows the median, upper quartile and lower quartile rank-one recognition rates for base classifiers used in the ensemble. The extent of the whiskers of the plot show the range of the values. We can see that as we increase the subspace size the boxes become more compact. This can be reflective of the "similar" nature of the subspaces and a lack of sufficient diversity. Similar results have been noted by [16, 19].

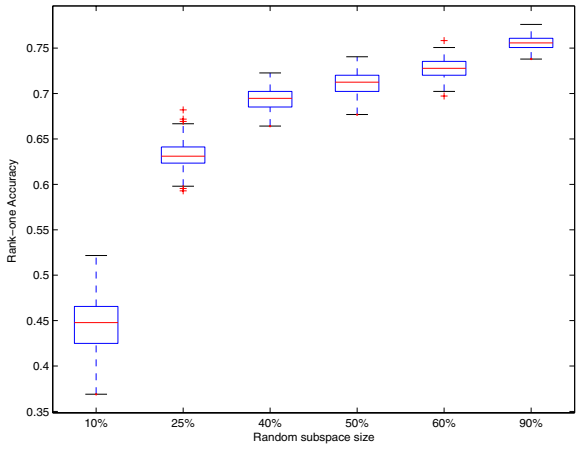

Figure 3. The accuracy spread among the classifiers learned on the random subspaces for *Data-1*.

We then evaluated the performance of random subspaces on *Data-2*; see Figure 4. As noted earlier, the probe and gallery images sets of *Data-2* were the same subjects as in the training, albeit with a temporal shift. Those images were taken at a different time than the training, with the probe image being the last available acquisition session. This is representative of the scenario when authentication of known subjects is required but images of those subjects are captured at a different time and (possibly) under different conditions. Thus requiring an improved generalization of the classifier(s). Since, the face space comprised of a large number of 1431 dimensions after PCA (the number of total images is 1432), we dropped 40% of dimensions [6] from

the back, as typically done. This tends to remove the low variance dimensions. We note that there were indeed significant improvements both in time and rank-one accuracy when using the "shrunk" face space. We applied random subspaces to this "shrunk" face space.

As is evident from the graphs, the random subspaces methodology consistently outperforms the single classifier. We see that 10% and 25% are usually better than the rest, with FB-LF as the exception. We are running additional experiments to analyze that effect. We are also running random subspace experiments on the entire face space that is all the 1431 dimensions to evaluate if random subspaces can alleviate the pruning requirement.

5 Conclusions

We evaluated the random subspaces methodology for the 2-D face recognition task. We ran multiple sized random subspaces with the nearest neighbor classifier. We showed that random subspaces is very competitive, and often outperforms a single nearest neighbor classifier learned on all the images. The random subspaces approach has the added advantage of requiring less careful tweaking (dropping of eigen vectors), which can be computationally expensive. We showed the usefulness of the ensembles in tackling the extreme components of dimensionality, time, and distribution variations in data. The generalization capacity of the ensembles is very relevant to the face recognition domain as the testing sets can be very different from the training set, and any overfitting can lead to reduced performance estimates. We observed using the box-plots that the individual classifiers are very weak, however an ensemble of the same provides an improvement in the accuracy. This diversity is pivotal to the success of the ensemble techniques. The ensembles usually increase the generalization, which can be helpful if there are different expressions and/or lighting conditions in the testing set.

The random subspace methodology can easily be set up in a distributed fashion, thus reducing the overall computational complexity. Since the subspaces are selected completely independent of each other, each (distributed) processor can look at a portion of the subspace and construct a classifier. In fact, we ran our experiments in a completely distributed fashion on a linux cluster.

Acknowledgments

This work is supported by National Science Foundation grant EIA 01-20839, Central Intelligence Agency, and Department of Justice grant 2004-DD-BX-1224. We would like to thank Jaesik Min for the help with data.

References

[1] K. Fukunaga and D. M. Hummels, "Bias of nearest neighbor error estimates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 103–112, 1987.

[2] K. Fukunaga and D. M. Hummels, "Bayes error estimation using Parzen and k-NN procedures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 634–643, 1987.

[3] I. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[5] G. Shakhnarovich and G. Moghaddam, *Handbook of Face Recognition*, ch. Face recognition in subspaces. Springer-Verlag, 2004.

[6] W. Yambor, B. Draper, and R. Beveridge, "Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures," *2nd Workshop on Empirical Evaluation in Computer Vision, Dublin, Ireland*, July 2000.

[7] Y. Hamamoto, S. Uchimara, and S. Tomita, "A bootstrap technique for nearest neighbor classifier design," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 73–79, 1997.

[8] T. K. Ho, "The random subspace method for constructing decision trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[9] N. V. Chawla and K. W. Bowyer, "Random subspaces and subsampling for face recognition," in *CVPR*, 2005.

[10] D. Beveridge and B. Draper, "Evaluation of face recognition algorithms (release version 4.0)." available at http://www.cs.colostate.edu/evalfacerec/index.html.

[11] X. Wang and X. Tang, "Random sampling LDA for face recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 259–265, 2004.

[12] V. Perlibakas, "Distance measures for pca-based face recognition," *Pattern Recognition Letters*, vol. 25, no. 6, pp. 711–724, 2004.

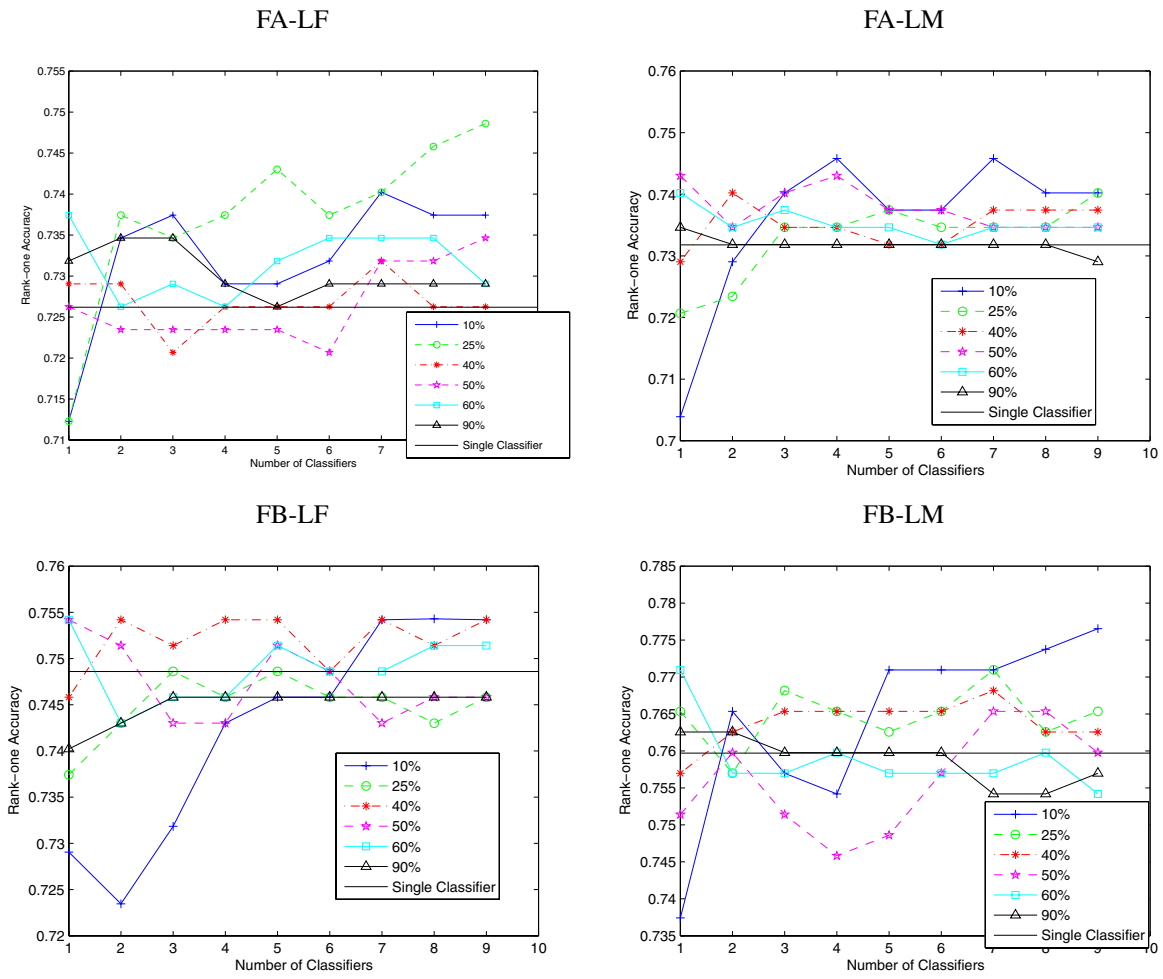

Figure 4. Results for *Data-2*.

[13] B. Draper and K. Baek, "Bagging in computer vision," in *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 144–149, 1998.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[15] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[16] T. Dietterich, "An empirical comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139 – 157, 2000.

[17] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Audio and Video Based Biometric Person Authentication*, pp. 44–51, 2003.

[18] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[19] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "Learning ensembles from bites: A scalable and accurate approach," *Journal of Machine Learning Research*, vol. 5, pp. 421–451, 2004.